WITHIN GROUP- VERSUS BETWEEN-GROUP CONSISTENCY: EXAMINING THE EFFECTIVENESS OF IRR TRAINING

David P. Baker
Casey Mulqueen
American Institutes for Research

R. Key Dismukes
NASA-Ames Research Center

ABSTRACT

Inter-rater reliability (IRR) training has been proposed as an effective strategy for training pilot instructors to accurately assess crew performance. This training usually takes place during a one-day workshop in which pilot instructors watch and assess the videotaped performance of several crews flying scenarios or their component event sets. While reasonable levels of inter-rater agreement have been reported for IRR training, these results are typically reported at the within–group level. At large air carriers, where pilot instructor/evaluators are trained in numerous workshops, between-group agreement is equally important. This paper explores the extent to which between-group differences exist across several IRR classes.

INTRODUCTION

As a result of the introduction of the Advanced Qualification Program (AQP), recent research in aviation psychology has focused on the process by which pilot instructors/evaluators (I/Es) are trained to assess aircrew performance and assign performance ratings during line operational evaluation (LOE) (Birnbach & Longride, 1993; George Mason University, 1996). Historically, airlines have employed what has been referred to as IRR (i.e., inter-rater reliability) training to meet this training need. IRR training usually consists of a one-day workshop in which I/Es practice assessing and rating the performance of crews flying LOE scenarios. For example, videotapes of crews flying specific LOE scenarios, or one of the component event sets, would be shown to a class of I/Es. These individuals would then independently rate each crew’s technical and crew resource management (CRM) performance on a grade sheet specifically designed for that LOE. During a class break, ratings are analyzed to determine the current level of calibration that exists within the group and areas where significant rating discrepancies exist. Upon reconvening the class, the results of these analyses are fed back to the workshop participants and rating discrepancies are discussed to reach consensus. Videotape of a different crew flying the same LOE is then rated to determine the level of calibration achieved (George Mason University, 1996).

Essentially, IRR training is a variation of frame-of-reference (FOR) training. In both cases, the primary goal is to train raters to a common frame of reference so that different I/Es will similarly assess LOE performance (e.g. an individual, aircrew, etc.). The fundamental difference between IRR training and FOR training, however, is the criteria to which I/Es are calibrated. Typically, IRR training involves providing feedback to I/Es on the extent to which I/Es agree with each other during training (i.e., a group standard). This process is accomplished by providing statistical information (e.g., interrater agreement, systematic differences, congruency, consistency, etc.) regarding the performance of each I/E relative to the rest of the group (Williams, Holt, & Boehm-Davis, 1997). FOR training, on the other hand, provides feedback to I/Es regarding their ratings in comparison to a gold standard. Gold standards are developed by task experts who carefully review the training videotapes and assign performance ratings on the LOE grade sheet. Consensus is then reached among task experts to arrive at the gold standard. These gold standards are believed to reflect the actual performance level displayed on the videotape (Baker, Swezey, & Dismukes, 1998; Sulsky & Balzer, 1988).

To date, a number of research studies have demonstrated the effectiveness of FOR training for reducing rating errors and increasing rating accuracy (Athey & McIntyre, 1987; McIntyre, Smith, & Hassett,1984; Pulakos, 1984, Pulakos, 1986; Woehr & Huffcutt, 1994). However, research has yet to demonstrate the effectiveness of IRR training. The basic question here is whether norming rater performance to a group standard (IRR training) is as effective as norming raters to a gold standard (FOR training). Testing the effectiveness of IRR training is important, because under IRR training it seems possible that I/Es could be shown to be calibrated within each IRR training class, but separate I/E IRR training classes might not be calibrated with each other. However, if IRR training was shown to be effective in promoting within- and between-group consistency, this training approach might be preferable, because it doesn't require the costly undertaking of developing gold standards for each IRR videotape.

The purpose of this study is to test the effectiveness of IRR training regarding its ability to improve within-group and between-group consistency. Regarding within-group consistency, we examined whether or not IRR training improves I/E rating accuracy by comparing I/E pre-training performance with post-training performance within several IRR training classes. Regarding between-group consistency, we compared the accuracy of I/E ratings across several IRR training classes. We focus on I/E rater accuracy in this study because high levels of accuracy are associated with high levels of rater agreement and consistency (Goldsmith & Johnson, in press). Based on the research reviewed, we hypothesized that IRR training should lead to an increase in I/E accuracy within an IRR training class, but comparable levels of accuracy may not be realized across IRR training classes.

METHOD

Participants

A total of 25 I/Es took part in IRR training, with classes 1 and 3 consisting of seven participants each and class 2 having eleven individuals. The participants were pilots from several different fleets at a major U.S. airline.

Rating Task

In the present investigation, participants viewed and rated two videotapes prior to training and one videotape after receiving feedback and discussing their pre-training ratings. Each tape displayed a different aircrew flying three event sets from the same LOE scenario. Crew 1 displayed average performance across the event sets, Crew 2 displayed below standard performance across the event sets, and Crew 3 displayed good performance across the events; Crews 1 and 2 served as the pre-training measure (i.e., pretest) and Crew 3 served as the post-training measure (posttest).

Ideally, we would have liked to have had I/Es rate more videos during the posttest or counterbalanced the order of the videotapes to rule out the possibility that improved performance on the post-training video was a function of the particular crew that was rated. However, this was not possible for several reasons. First, with respect to including additional videotapes for the posttest, we found that videotapes of aircrews flying LOE scenarios are hard to collect given the highly confidential nature of this information. To date, the airline has done a tremendous job collecting such information for IRR training. Nonetheless, only a very few videotapes were available in the airline's videotape library, and of those videos, the ones selected for IRR training were the most current and relevant (i.e., the tapes displayed aircrews flying the LOE to be used by the airline in the future). Furthermore, the addition of videotapes to be rated during the posttest would have significantly increased the length of IRR training, and each IRR class was limited to one full day. Rating each crew required approximately an hour of participant time. Therefore, the pretest and the posttest combined required three hours. This left one-half day to conduct the training. Second, data for this study were gleaned from archival data sources. Therefore, we were unable to dictate which videotapes were included and the order in which they were used during IRR training

Participants in IRR training rated the performance of each crew on the three LOE event sets using the airline's LOE grade sheets. These grade sheets consist of three components. First, I/Es are required to assess the extent to which specific CRM behaviors are observed, partially observed, or fully observed. Second, I/Es evaluate a series of specific technical skills on a four-point scale where 1=repeat, 2=debriefed, 3=standard, and 4=excellent. Finally, I/Es evaluate the overall performance of the crew for each scenario event set. Grades are made on the four-point rating scale regarding the crew's overall CRM performance, overall technical performance, and individual performance of each crewmember (i.e., Captain and First Officer).

Dependent Measures

Gold standards were established for each IRR videotape by convening an independent panel of experts to review and evaluate each crew's performance on the LOE event sets. Gold standards were established for the CRM behavior ratings, the technical skill ratings, and the overall CRM, technical and crewmember performance ratings that are recorded for each event set on the LOE grade sheet. This process mirrored the procedure for establishing gold standards described by Baker et al. (1998). Gold standards were developed for both the pre-training and post-training videotapes.

Distance accuracy (DA), the dependent measure in this investigation, provides a measure of agreement between each I/E's ratings and the gold standard ratings. DA is calculated by determining the absolute value of the deviation of each I/E's rating from the gold standard. Average DA scores were then calculated for (1) CRM behaviors, (2) Technical skills, (3) Overall Event Set grades (CRM and Technical), (4) Overall grades for the Captain and First Officer. DA scores for the two videotapes that were rated prior to training (i.e., Crews 1 and 2) were averaged to create a measure of pre-training performance, and DA scores for Crew 3 were used as a measure of post-training performance.

IRR Training

IRR training was conducted in the same fashion for each of the three training classes. First, descriptive information was presented to I/Es regarding the LOE grade sheet and scenario event sets to be evaluated. Second, each I/E watched and rated Crews 1 and 2 on the three LOE event sets. Videotapes were presented in such a fashion that I/Es watched and independently evaluated the performance of Crew 1 on Event Set 1 and then watched and evaluated Crew 2, Event Set 1 and so on for all three event sets. These ratings comprised the pre-training assessment and were used as a basis for discussion during the next phase of IRR training. During a class break, data from the pre-training phase were analyzed to determine the levels of consistency and agreement that existed across I/Es within the class. Upon reconvening the class, these data were fed back to I/E trainees and any rating discrepancies were discussed. For example, the IRR facilitator would present the results of the agreement analysis for Crew 1, Event Set 1. Items (e.g., CRM behaviors, technical skills, etc.) on which I/Es were in significant disagreement were identified and discussed. This discussion examined why different I/Es gave the same crew different performance ratings. The goal of this discussion was to develop common standards within the I/E group. To the extent it was possible, all discrepant ratings identified through data analysis were discussed. Once the feedback and discussion phase was competed, I/Es rated the performance of the crew on the post-training videotape.

RESULTS

Pre-Training Class Differences

Because random assignment to training classes was not possible, it was necessary to show that the training classes were reasonably similar on the pre-training measure. Based on the method that the airline used to assign pilots to each of the classes, it was reasonable to assume that no group differences between the three classes on the pre-training measure existed. Table 1 provides the DA means and standard deviations for each class for each type of rating made for the pre-training videotapes. For each of the classes, disagreement with the gold standards becomes somewhat more pronounced for the Event Set and Overall grades. However, none of the classes were significantly different regarding their ratings (CRM, technical, etc.) of the pre-training videotapes.

Table 1. DA means (M) and standard deviations (SD) for the three IRR classes for the pre-training videotapes.

	Class 1		Class 2		Class 3	
	M	SD	M	SD	M	SD
CRM	.46	.09	.40	.09	.44	.11
Technical	.45	.11	.45	.14	.47	.13
Event Set	.62	.16	.66	.18	.75	.28
Overall	.58	.18	.79	.18	.58	.23

Note: N=7 (Class 1); N=11 (Class 2); N=7 (Class 3).

IRR Training Effectiveness

In order to test the effectiveness of IRR training, pre-training measures of DA were compared to the post-training measures of DA using paired-samples t-tests. These comparisons were made for each IRR training class and based on the combined data from the three IRR classes. Table 2 reports the mean pre-training and post-training DA values for each class. It can be seen that, generally speaking, there is better agreement (lower DA) for the post-training measures indicating that IRR training was effective, with several means significantly different at $p < .05$.

Table 3 reports the mean pre-training and post-training DA values based on the combined data from the three IRR classes. Referring to Table 3, significant mean differences are observed for all LOE grade sheet items. These results provide additional and stronger support for the conclusion that IRR training significantly improved pilot I/E rating accuracy.

Table 2. Pre-training and post-training DA by IRR class.

	Pre-training		Post-training	
	M	SD	M	SD
Class 1				
CRM	.46	.09	.36	.12
Technical	.45	.11	.29	.15
Event Set	.62	.16	.38	.25
Overall	.58^{a}	.18	.33^{a}	.19
Class 2				
CRM	.40	.09	.36	.08
Technical	.45^{b}	.14	.27^{b}	.20
Event Set	.66	.18	.50	.17
Overall	.79c	.18	.40c	.23
Class 3				
CRM	.44	.11	.31	.13
Technical	.47^{d}	.13	.14^{d}	.09
Event Set	.75^{e}	.28	.15^{e}	.06
Overall	.58^{f}	.23	.04^{f}	.12

Note: N=7 (class 1); N=11 (Class 2); N=7 (Class 3). Means denoted by the same letter are significantly different at p<.05.

Table 3. Pre-training and post-training DA.

	Pre-training		Post-training	
	M	SD	M	SD
CRM	.43^{a}	.09	.34^{a}	.10
Technical	.46^{b}	.13	.24^{b}	.17
Event Set	.67^{c}	.20	.37^{c}	.23
Overall	.67^{d}	.21	.28^{d}	.24

Note: N=25. Means denoted by the same letter are significantly different at p<.05.

Post-Training Class Differences

In addition to determining whether or not IRR training produced gains in post-training accuracy, we examined whether or not these gain were similar across the three training classes. A between-subjects ANOVA on the post-training data found a main effect for class for Event Set grades (F (2, 22) = 8.718, p < .05) and for Overall grades (F (2, 22) = 7.534, p < .05). In other words, although classes were similarly accurate prior to delivery of IRR, and IRR training improved I/E rating accuracy, it did so somewhat differently across the three IRR classes. Post hoc tests indicated that for the event set and overall grades that Class 3 was significantly more accurate after IRR training than Classes 1 or 2 (refer to Table 4).

Table 4. DA means (M) and standard deviations (SD) for the three classes for the post-training videotapes.

	Class 1		Class 2		Class 3	
	M	SD	M	SD	M	SD
CRM	.36	.12	.36	.08	.31	.09
Technical	.29	.15	.27	.20	.14	.09
Event Set	.38^{a}	.25	.50^{b}	.17	.15^{ab}	.06
Overall	.33^{c}	.19	.40^{d}	.23	.05^{cd}	.12

Note: N=25. Means denoted by the same letter are significantly different at p<.05.

DISCUSSION

The results of this investigation provide preliminary support for the hypothesis that IRR training can lead to increases in I/E accuracy within separate IRR training classes, although comparable levels of accuracy may not be realized across IRR training classes. Significant gains in rating accuracy were observed within each of three IRR training classes, but similar gains in rating accuracy were not observed across these IRR training classes. This result occurred despite standardized IRR training procedures and the same IRR facilitator conducting each class. The sole variation rested in the fact that feedback in each class was based on group standards as opposed to gold standards. While group standards appear to lead to greater within-group consistency they do not necessarily produce greater between-group consistency. This issue is particularly important for large air carriers who must train their pilot I/Es in separate IRR training classes, because the goal of IRR training is to produce accurate and consistent LOE assessment throughout the I/E population. If similar levels of accuracy and consistency are not achieved across I/Es then a crew's performance in an LOE will depend more upon the I/E conducting the assessment than the crew's performance.

To emphasize this point further, Table 5 presents the post-training mean ratings for each of the three IRR training classes. Referring to Table 5, slight differences are observed across each class. Class 2 gives higher ratings to Crew 3 while Class 3 gives lower ratings to Crew 3. These differences are particularly pronounced for the event set and overall ratings. Here, the average rating for Class 2 is almost a .5 scale point higher than Class 3 (on a 4-point rating scale). When combined with the fact that Class 3 produced the most accurate assessment of Crew 3 (refer to Table 4), it can be concluded the Class 2 participants normed to a more lenient group standard.

Table 5. Post-training means (M) and standard deviations (SD) for each IRR class.

	Post-training	
	M	SD
Class 1		
CRM	2.64	.46
Technical	3.16	.43
Event Set	3.05	.57
Overall	3.10	.48
Class 2		
CRM	2.91	.20
Technical	3.25	.43
Event Set	3.44	.45
Overall	3.37	.47
Class 3		
CRM	2.83	.26
Technical	3.01	.29
Event Set	2.95	.29
Overall	2.95	.13

Note: N=7 (class 1); N=11 (Class 2); N=7 (Class 3).

The results of this study suggest that gold standards training may be a more effective strategy than IRR for training pilot I/E rating accuracy. While gold standards training may be more costly and time consuming to develop, this approach relies upon standardized feedback (i.e., gold standard) as oppose to individual rating norms that are developed within each IRR class (i.e., group standards). In fact, the investigation presented here represents the first stage of a larger investigation that will compare the efficacy of IRR and gold standards training.

Limitations and Future Research

As noted previously, random assignment of I/Es to the training classes was not possible for this study. In a strict experimental control sense, this can be viewed as a weakness. However, analysis of class differences prior to training indicated that pre-existing differences among pilots was not an issue. A larger number of trainees per class would have added more power to our analyses. However once again, the experimental rigor of the study is limited by the availability of I/Es at the airline in training. We believe that our limited sample size is offset by the fact that the study was conducted with actual pilot I/Es engaged in actual IRR training classes, adding to more reliable generalizability of results. Regarding the training materials, the inclusion of an additional post-training videotape would have helped to stabilize the reliability of results, as would the counterbalancing of video presentations between classes. Once again, the constraints of conducting field, as opposed to laboratory research, made this impossible.

The current study investigated only the effectiveness of IRR training, and is a first step in the process of comparing the utility of IRR to gold standards training. This initial phase of investigation has been useful for determining the baseline effectiveness of IRR. Inclusion of the gold standards and control group conditions will establish the relative effectiveness of each training method.

AUTHOR'S NOTE

This research was supported by a grant from the NASA Ames Research Center. The views presented in this paper are those of the author(s) and should not be construed as an official NASA position, policy or decision, unless so designated by other official document. Address all correspondence to David P. Baker, American Institutes for Research, 3333 K Street, NW, Washington, DC 20007 or e-mail dbaker@air.org.

REFERENCES

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels of processing theory and social facilitation theory perspectives. *Journal of Applied Psychology*, *72*, 567-572.

Baker, D. P., Swezey, R. W., & Dismukes, R. K. (1998). *A methodology for developing gold standards for rater training videotapes*. Washington, DC: Federal Aviation Administration, Office of the Chief Scientific and Technical Advisor for Human Factors.

Birnbach, R. A., & Longridge, T. M. (1993). The regulatory perspective (pp. 263-281). In E. L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management*. New York: Academic Press.

George Mason University. (1996). *Developing and evaluating CRM procedures for a regional air carrier, phase I report*. Washington, D.C.: Federal Aviation Administration.

Goldsmith, T. E., & Johnson, P. J. (in press). Assessing and improving evaluation of aircrew performance. *International Journal of Aviation Psychology*.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology*, *69*, 147-156.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69,* 581-588.

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. *Organizational Behavior and Human Decision Processes, 38*, 76-91.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73,* 497-506.

Williams, D., Holt, R., & Boehm-Davis, D. (1997). Training for inter-rater reliability: Baselines and benchmarks. *Proceedings of the 9th International Symposium on Aviation Psychology*, 514-520.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, *67*, 189-205.